# サポートベクターマシンによる日本語長文の短文分割

根岸 知弘　高木 一幸　尾関 和彦

電気通信大学大学院 電気通信学研究科
{esnegish,takagi,ozeki}@ice.uec.ac.jp

## 1 はじめに

日本語に現れる長文はその構造に多数の接続節を含むため，係り受け構造の曖昧さが飛躍的に増加する．このため，長文をそのまま係り受け解析することは非常に困難である [1]．そのために係り受け解析の補助手段として，長文をより単純な構造の短文へと分割することが研究されている [1, 2, 3].

長文を短文へと分割する手法として，長文を短文へと分割する分割パターンを作成し，入力文とのパターンマッチングにより分割点を推定する手法や [2, 3]，ヒューリスティックなスコアを用いた分割規則によって分割点を推定する手法がある [4]. これらの手法はそれぞれ有効であることが示されているが，分割パターンや分割規則を人手で作成する必要がある．このため一貫性，網羅性の点などに問題があり，複雑な言語現象を十分に捉えきれているとは言えない．

このことから，短文への分割点を推定するためのパターンを人手で作成するのではなく，決定木を用いることでコーパスから自動的に獲得する手法が提案されている [1]．この手法により，コーパスから短文分割パターンを自動的に獲得できることが示されている．そこで本研究では学習モデルとしてサポートベクターマシン (SVM) を用い，コーパスから分割点を自動的に推定する方法を提案する．

## 2 サポートベクターマシン

サポートベクターマシン，SVM は 1995 年に，AT&T の V.Vapnik 等 [5] によって提案された $n$ 次元 Euclid 空間上の識別平面を用いた線形 2 値識別器である．

また線形可分でないデータに対しても，カーネル関数を用いて識別を行なうことができる[1]．カーネル関数は線形可分ではない学習データを，写像 $\phi$ を用いてより高次元な線形空間に写像することで識別を行なう．カーネル関数として，例えば

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^d \quad (1)$$

がある．式 (1) は多項式カーネル関数と呼ばれ，学習データにおける $d$ 個の素性の組み合わせを考慮した学習が可能であることが知られている．

[1] 線形可分でないデータに対する手法として他に，学習における多少の識別誤りを許すソフトマージン手法がある．

## 3 SVM を用いた長文の自動短文分割

### 3.1 短文分割点の定義

文を短文に分割するためには，「短文」の定義を明確に定める必要がある．本研究では「並列節」に着目することで，係り受け解析の補助手段としての短文を定義する．

日本語において述語を中心としたまとまりを「節」と呼び，そして二つ以上の節からなる文を複文と呼ぶ．複文は文末の述語を中心とした「主節」と，それ以外の節である「接続節」に分けられる．この「接続節」は「主節」との関係から「並列節」と「従属節」にさらに分けられる [6].

係り受け構造において「主節」と並列するということは，「並列節」の末尾の文節が文末の述語に係ることを意味する．つまりこの「並列節」を見つけることで係り受け解析を容易にすることができる．そこで短文分割の問題を文の「並列節」を検出する問題とし，この「並列節」の直後を「短文分割点」と定義する [1]．本研究では短文分割点として，「並列節」の末尾の文節が用言文節であるものに着目する．

以上から短文分割点は文末文節に係る文節を検出することで発見できる．しかし文末の文節に係る文節の中には主節に含まれるものがある．このような文節を短文の末尾とすることは適切ではない．そこで主節の範囲の認定を次のように行うことで短文分割点を決定する．文末文節との間に，

1. 用言文節
2. 副詞

3. 係助詞「は」,「も」
4. 格助詞「が」,「を」,「に」

などで終わる文節の中の少なくとも一つが存在する用言文節の中で，最も文末に近いものを検出する．この用言文節の次の文節から文末文節を主節とする．

以上のような短文分割点の定義から，コーパスにより事前に係り受け構造のラベルが付与されていれば，機械的に短文分割点を検出できる．

### 3.1.1 データの定義

**要素文節**

文中の文節には短文分割点を推定する上で重要な働きをするものと，あまり重要な働きをしないものがあると考えられる．節を構成するためには用言文節が必要であり，末尾が係助詞「は」,「も」，格助詞「が」である文節も重要な働きをすると考えられる [1]．このような文節を要素文節とし，その属性を学習に用いる．

**要素文節の属性**

要素文節の属性は着目した文節を構成する形態素の表層情報を利用する．本研究における着目した文節の主辞品詞は，基本的に自立語[2]に属する品詞分類を持つ形態素の中で，最も文節末尾に近い形態素の品詞とする．ただ接尾辞を伴う場合には，接尾辞に属する形態素の中で，最も文節末尾に近い形態素の品詞細分類により主辞品詞を定める[3]．

(a) 接続属性

接続属性の属性値を上記で定めた主辞品詞により表 1 のように定める．主辞欄の「体言判定」は体言に判定詞が付属したものを表し，形容動詞は主辞品詞が形容詞で活用語型がナ形容詞であるものとする．また「用言」は「動詞」,「形容詞」,「形容動詞」,「体言判定」を含む．

「A 類」,「B 類」,「C 類」とは，接続節の末尾の形態による接続節の独立性の変化に着目し，接続節を分類したものである [6]．これらは「C 類」,「B 類」,「A 類」の順に独立性が下がる[4]．また属性「て」はこの文献 [6] においては「A 類」,「B 類」,「C 類」の複数の類に属するので，独立した属性「て」として分類する [1]．

形容詞と形容動詞の「連用形」は述語の修飾語として働くことができるため [7]，動詞の「連用形」とは働きが異なることが予想される．そこで「形連用」,「形動連用」に分類する [1]．

(b) スコープ属性

用言文節に形式名詞「こと」や引用助詞「と」などが含まれていると，それより前にある文節がそれを飛び越して文末の文節に係ることは少ない [3]．

このような現象を利用するため「スコープ属性」を設ける [1]．その属性値は，文節に引用助詞「と」，助動詞「ようだ」，形式名詞，副詞的名詞「よう」,「ところ」，副助詞「など」，格助詞+接続助詞「との」のいずれかが含まれる場合には「スコープ」，ない場合には「NULL」と定める．

(c) 読点属性

読点を伴う文節はそこに構文的な区切りが存在し，離れた文節に係ることを示すと考えられる．そこで読点による接続節の独立性の変化を利用するために「読点」属性を設ける．その属性値は文節末尾に読点「、」があれば「読点」，ない場合には「NULL」と定める [1]．

**短文分割点候補**

要素文節において，接続属性の値が「動連用」,「て」,「A 類」,「B 類」,「C 類」,「形動連用」,「形連用」のいずれかである場合に，その直後が短文分割点になる可能性がある [1]．そこでこれらの文節のうち主節の範囲に含まれない文節を分割文節候補とし，その直後を分割点候補とする．

**SVM の入力データ**

分割点を推定する際に，分割文節候補の属性値が重要である．そして分割点候補が分割点になる

[2] 名詞，動詞，形容詞，副詞，連体詞，指示詞，接続詞，感動詞．

[3] 名詞性接尾辞 → 名詞，動詞性接尾辞 → 動詞，形容詞性接尾辞 → 形容詞．

[4] 「A 類」の接続節は「B 類」の接続節に含まれることが可能で，「A 類」,「B 類」は「C 類」に含まれることが可能 [6]．

表 1: 要素文節の接続属性

| 属性値 | 主辞品詞 | 文節末尾の形態素の品詞, 表記, 活用形 |
|---|---|---|
| 動連用 | 動詞 | 連用形 |
| て | 動詞, 体言判定 | 連用テ形 |
| A 類 | 用言 | 接続助詞: 「ながら」,「つつ」 連用タリ形 |
| B 類 | 用言 | 接続助詞: 「と」,「まで」,「なら」 助動詞: 「ので」,「のに」,「ず」, 「ないで」, 条件形 動詞: 連用テ形+副助詞「も」, 条件形 |
| C 類 | 用言 | 接続助詞: 「が」,「から」,「けれど」, 「けれども」,「し」 |
| 形動連用 | 形容動詞 | 連用形 |
| 形連用 | 形容詞 | 連用形 |
| 動連体 | 動詞, 体言判定 | 連体形 |
| 用言 | 用言 | A 類, B 類, C 類以外の 接続助詞, 接続助詞以外の助詞 |
| 形連体 | 形容詞, 形容動詞 | 連体形 |
| は | 体言 | 副助詞「は」 |
| も | 体言 | 副助詞「も」 |
| が | 体言 | 格助詞「が」 |
| 基本形 | 用言 | 基本形 |
| 終止形 | 用言 | 基本形 (タ形)+句点 |

か否かは，分割文節候補が分割文節以降のどの文節に係るかによって決まる．つまり分割文節候補より後に現れる要素文節も重要な役割を持つと考えられる [1]．そこで本研究では分割文節候補と分割文節候補以降に現れる要素文節の列を，一つの分割候補要素文節列として SVM の入力データとする．学習データと評価データには分割点候補が分割点であるか否かによってそれぞれ正例 (YES), 負例 (NO) のラベルを付与する．

# 4 実験

## 4.1 京都大学テキストコーパス

本研究では言語データベースとして，京都大学テキストコーパス (version3.0)[8] を利用した．本研究ではデータを作成する際に，コーパスから以下の条件に該当する文は取り除いた．

1. 形態素の属性項目に情報の欠如がある文
2. 単文節である文
3. 一文の中で複数の句点「。」がある文
4. 自立語が欠けた文節を含む文
5. 未定義語を含む文
6. 係り受け関係の表示が誤っている文

## 4.2 実験データの作成

京大コーパスにおいて助詞「と」はほとんどが格助詞に分類されている[5]．このことから格助詞に分類される「と」で，その文節が用言に係るものを引用助詞として扱う．学習データとして約 4000 文のデータから短文分割点候補を検出し，分割候補要素文節列データを作成した．また評価データとして約 2000 文から同様にデータを作成した．

## 4.3 実験における精度の評価尺度

SVM の分割精度を評価する尺度として以下のものを採用する [9].

$$\text{適合率} = \frac{\text{正しく “}YES\text{“ と判定されたデータ数}}{\text{“}YES\text{“ と判定されたデータ数}}$$

$$\text{再現率} = \frac{\text{正しく “}YES\text{“ と判定されたデータ数}}{\text{“}YES\text{“ のラベルが付与されたデータ数}}$$

$$\text{文正解率} = \frac{\text{完全に正しく分割されたデータ数}}{\text{評価文数}}$$

また適合率と再現率を一つの尺度として評価するために，それらの調和平均である F 値 [10] も用いる．

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

## 4.4 実験結果

京都大学テキストコーパス [8] に対し，上記の方法で SVM$^{light}$[11] による短文への分割点推定実験を行った．学習データにおいて分割できた文数が 2038 文，分割候補要素文節列データが 3365 個，評価データでは分割できた文数が 1078 文，分割候補要素文節列データが 1928 個得られた．

これらのデータを多項式カーネル関数を用いて[6]短文分割点の推定実験を行った結果，表 2 のような結果が得られた．

誤りが多かった分割点候補要素文節列のパターンは次のようなパターンである．ここで短文分割点を「||」と表す．

< (動連用 NULL 読点) || (基本形 NULL NULL) >

[5]格助詞以外では接続助詞があるが格助詞に対する割合は 0.001%程度.

[6]次元数 $d$ は 3 とする.

表 2: polynomial カーネル関数を用いた短文分割

| $C$[7] | 適合率 | 再現率 | F 値 | 文正解率 |
|---|---|---|---|---|
| 0.0027 | 76.75% | 83.65% | 80.05% | 71.89% |
| 1.0 | 74.65% | 80.52% | 77.48% | 69.48% |
| 2.0 | 74.80% | 80.40% | 77.50% | 69.48% |
| 3.0 | 74.80% | 80.40% | 77.50% | 69.48% |

SVM ではこのパターンを，短文分割点ではない (NO) ラベルを付与されているのに，短文分割点である (YES) と誤って識別している．これは他のデータにおいて次の例文 1,2 が現れるためである．例文 1,2 は要素文節の定義により下線部を上記のパターンとして解釈する，例文における文節と要素文節の対応を表 3 に示す．

表 3: 例文における下線部の文節と要素文節の対応

| | 分割文節候補 (動連用 NULL 読点) | 要素文節 (基本形 NULL NULL) |
|---|---|---|
| 例文 1 | すすめ,[10] | 語る[15] |
| 例文 2 | とどまり,[6] | 求める[8] |

[例文 1 ] ライカ人類の[1]常識と[2]非常識を[3]説き,[4]三万円コースから[5]始まる[6]「予算別買い物ガイド」で[7]中古品の[8]個人輸入を[9]すすめ,[10] || 最後に[11]「二〇二四年フォトキナ現地レポート」で[12]M 3 再登場の[13]夢を[14]語る[15]三百九頁.[16]

[例文 2 ] 赤字額は[1]政府が[2]見積もった[3]国内総生産の[4]約七・八%に[5] とどまり,[6]|| 緊縮財政を[7]求める[8]国際通貨基金の[9]要請を[10]辛うじて[11]クリア.[12]

例文 1,2 において共通する点は文末の体言である．例文 2 はコーパス上で下線部先頭の分割文節候補の連用形が文末の体言を修飾し，連用形の直後を短文分割点とする．これに対し，例文 1 は下線部先頭の分割文節候補の連用形が文末の体言を修飾せず，連用形の直後を短文分割点としない．このような動詞の連用形に修飾されうる体言の情報抽出の不足が精度の低下の原因の 1 つとなっている．

## 5 むすび

本研究では表層情報による短文分割点の推定に対して，SVM の手法を提案した．京都大学テキストコーパスを用いて提案した手法による実験を行った．その結果，SVM による短文分割点推定において，適合率 77%，再現率 84%，文正解率 72%が得られた．これにより SVM による短文分割点に関するコーパスの表層情報を文全体から生成し，これを学習することによって，短文分割点を自動的に推定できることがわかった．

しかし，まだ短文分割点に関するコーパスの表層情報を十分に抽出しているとは言えず，情報の不足による識別誤りが精度の低下の大きな要因となっている．また識別誤りを起こした分割候補文節を先頭とする要素文節列のパターンは多岐にわたっている．

今後の課題としては，短文分割点として副詞に相当する役割を持つ名詞などの属性の追加や，表層情報を詳細化し，その情報の抽出の精度をより向上させることが挙げられる．また短文分割点の推定結果を係り受け解析に利用し，係り受け解析の精度に対する効果を調べる必要がある．

[7] $C$ はソフトマージン手法における識別誤りの度合いを示す変数.

## 参考文献


[1] 張 玉潔, 尾関 和彦, "決定木による日本語長文の短文分割, " 自然言語処理, Vol. 7, No. 1, pp. 13-30,2000

[2] 金 淵培, 江原 暉将, "日英機械翻訳のための日本語長文自動短文分割と主語の補完, "情報処理学会論文誌, 35.No.6, 1018-1028, 1994

[3] 黒橋 禎夫, 長尾 真, "並列構造の検出に基づく長い日本語文の構文解析, " 自然言語処理, Vol. 1, No. 1, 35-57, 1994

[4] 武石 英二, 林 良彦, "接続構造解析に基づく日本語複文の分割, " 情報処理学会論文誌, 33.No.5, 652-663, 1992

[5] C.Cortes and V.Vapnik. "Support Vector Networks. Machine Learning, " Vol.20, pp.273-297, 1995

[6] 南 不二男, "現代日本語の構造, " 大修館書店, 1974

[7] 益岡 隆志, 田窪 行則, "基礎日本語文法, " くろしお出版, 1974

[8] 京都大学テキストコーパス Version3.0, 2000 http://www.pine.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

[9] 永田 昌明, "単語頻度の再推定による自己組織化単語分割, " 情報処理学会研究報告, NL-121, 9-16, 1997

[10] D.Hindle and M.Rooth, "Structural ambiguity and lexical relations Computational Linguistics, " 19(1), 103-120, 1993

[11] Thorsten Joachims, "SVM$^{light}$ version:5.00, " 2002 http://svmlight.joachims.org/